**Title:** Preliminary proposal to add a new provisional kIDS property (Unihan)

Author: Ken Lunde & John H. Jenkins

**Date:** 2021-06-30

The purpose of this document, which is a preliminary proposal to add a new provisional Unihan database property, *kIDS*, is threefold:

- 1. Outline the standardization timeline
- 2. Identify any barriers early on in the process through constructive and meaningful feedback from both the UTC and IRG
- 3. Solicit help in collecting ideograph components for use in IDSes

IDS is an abbreviation for *Ideographic Description Sequences*, which is documented in Section 18.2, *Ideographic Description Characters*, of the Core Specification.

The standardization timeline has two targets: Unicode Version 15.0 (2022) and Unicode Version 16.0 (2023).

## Unicode Version 15.0 (2022)

The targets for Unicode Version 15.0 are to:

- 1. Encode four or five new Ideographic Description Characters (IDCs)
- 2. Encode a modest number of ideograph components as new CJK Unified Ideographs for use in IDSes

This then sets the stage for adding the provisional *kIDS* property in the subsequent version of the Unicode Standard.

## **New Ideographic Description Characters**

Four new IDCs were most recently proposed in L2/18-012 (aka IRG N2273) as shown in the table below:

IDC	Туре	Character Name
[]	Binary	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM RIGHT
	Binary	IDEOGRAPHIC DESCRIPTION CHARACTER SURROUND FROM LOWER RIGHT
$\longleftrightarrow$	Unary	IDEOGRAPHIC DESCRIPTION CHARACTER HORIZONTAL REFLECTION
$\bigcirc$	Unary	IDEOGRAPHIC DESCRIPTION CHARACTER ONE HUNDRED EIGHTY DEGREE ROTATION

The Ideographic Description Characters block, which is the most appropriate block for encoding these four new IDCs, has exactly four available code points: U+2FFC through U+2FFF. The proposal will recommend encoding these new IDCs using these particular code points.

There is a fifth new IDC that is being considered whose recommended name is IDEOGRAPHIC DESCRIPTION CHARACTER STROKE SUBTRACTION. This new IDC is binary, and is followed by two components:

- 1. An ideograph component
- 2. A CJK stroke from the CJK Strokes block that is omitted from the ideograph component, or multiple CJK strokes if an IDC is used

Below are examples that use U+2296 ⊖ CIRCLED MINUS as a placeholder for the actual IDC:

- The IDSes for U+2002A 其 and U+2002B 其 are difficult to represent with existing ideograph components, but could be represented as ⊖其 / and ⊖其 \, respectively.
- The IDS for U+2CEBB  $\overline{s}$  is also difficult to represent with existing ideograph components, but could be represented as  $\Theta$

A counter-example for the first example above would be to instead encode the common ideograph component of U+5176  $\pm$ , U+2002A  $\pm$ , and U+2002B  $\pm$  as a new ideograph component, but that handles only this particular case.

## **New Ideograph Components**

A non-trivial number of CJK Unified Ideographs include components that cannot easily be represented in IDSes, and a modest number of additional ideograph components—encoded as new CJK Unified Ideographs—would serve to improve the IDSes of such ideographs.

As a mere sample, the table below illustrates the numbered ideograph components based on the 2021-05-26 version of IDS.TXT, and provides examples of ideographs whose IDSes include them (ideographs for which there is not an available font are shown as a placeholder glyph followed by their IDSes shown in green and enclosed in red brackets):

Syntax	Component	Examples
{01}		<u>馬</u> 焉爲勵
{02}	伍	免晚象劖劖鳨ګ舄絻睌爮敟芻齹色
{03}	负	复 與 檢 复 做 多 換 換 換 換 換 換 換 換 換 換 換 換 換 換 換 換 換 換
{04}	भार	<b>紫</b> 紫
{05}	F	叚罾屗跂麫段鰤段段假
{06}	F	段韈茛幵緞葮熋鰕

Syntax	Component	Examples
{07}	重	專惠懷壽惠慰捷臺脯電則惠專繼進鼉街連
{80}	七	扥 <b>酰</b>
{09}	7	新假催光午따 的 對
{10}	凡	嬴巩羸羸鸁鸁鸁鸁鸁鸁鸁羸羸羸羸羸羸羸羸羸 鸁鸁鸁鸁鸁鸁鸁
{11}	><	兆函塁摂楽氺渋率疂脊軣雨鰥艰脫半堀賡挑葢蕸麔 艦鰥鞯討湍兆鎟剟眘炎贵蹟閴
{12}		龍龍喧種龎
{13}		卢睿卢定府
{14}	<u></u>	侯候叏夬等命保候具愛宾楙発通
{15}	氏	旅玈祣ҡҞ胨厎低嫙搲疷玈震骶點延返
{16}	BBL	巤鬛鼠鼢獵攨遠鬛瀘稵鏥鬛
{17}	良	剆勆朗欴郎郒剆飡訢蘅缻訮誁鵖鹍即竛
{18}	艮	匓即廄既鄉鵙討欴螂髸愍節
{19}	更	更更使理)更更使理
{20}	5	埓將搖望炙然瑤謠遙凰滛禽
{21}	夕又	祭詧豋攆樥雞尞藿唵粲鰲

Syntax	Component	Examples
{22}	与	与写复号音號
{23}	Ħ	與璺叟魁臂臋與黰學罍鬱
{24}	K	<b></b>
{25}	LX.	丧喪展畏辰長丧畏衮表 是 浪 異 級 累 泉 虎 液 核 長 畏 衮 丧 坂 冢 衮 辰
{26}	流	亚齊
{27}	4	丝没支衫
{28}	7	幾畿幾畿幾幾天武武
{29}	丝丝	幾畿幾畿幾幾幾
{30}	Ħ	<b></b>
{31}	幺	卤囪圙馸頋圛馸曥
{32}	仄	<b></b>
{33}	巾巾	雋 <u>假</u> 觀 明 图 [ [ ] ] 雅 {33} ]
{35}	凡	冺 <u> </u>
{36}	戸	憂戞忢뤗募臱瓇獶薆聂犥瓇媷囁蚤
{39}		兜卬兒兒遛留卬銞屲叽
{40}	异	亞亞猫戶

Syntax	Component	Examples
{41}	$\exists$	億 会 妇 旋 寻 尋 吊 归 当 彗 急 扫 曜 榋 灵 雪 尋 捷 龝 卪 侯 侵 易 阳 曼 良 啑 遐 竭 嫁 露 寤 尋 美 屆 帮 印 書 起 录 曩 曩 蒙 霧 襲 艮 意 報 報 設 設 貴 書 程 樏 散 浆 累 書 冒 禮 禄 叙 第 書 智 禮 報 設 惠 景 電 表 書 表 書 表 書 表 書 表 書 表 書 表 書 表 書 表 書 表
{42}	∄.	
{43}	<b>⇒</b>	争事事使兼妻媫尋尹帚庚彗彛急灵爭疌秉粛肀隶雪 灵捷哥事侵燕荛录受嗝啑塌寣烡炭彛孚ಠ捏摓鼓散 書東棲菜菜椗淒濅爭寁淾康睫禝雛纂耨屐臛蒂蓬 蔧蓤蜺誱逑鍉鵑事捿孚緑譩禄傔赓釹掲覯謙昌②[[[]] 性{43}]②[[[]]{43}者]②[[[]][[]][[43]二水]
{47}	鳥	島梟裊鳥鳬鳧裊濠追梟鳧鷦鳥島邊
{49}	IJ	帅师帰归班疈辨所坚爽誹?[[[[{49}无]]]
{50}	コ	兜段囙尸巨彐戶目禮元兒兒兒娘愛問段榮疆我或骸 斗 最 卑 撒 混 捏 提 报 嚴 馨 馨 惠 盝 且
{51}	馬	馬凴馬馬馬馬馬馬
{52}	马	马
{53}	ウ	<b>凫岛枭袅鸟岛</b>
{54}	島	鳥島
{55}	广	馬慶鹿

Syntax	Component	Examples
{56}	<del>===</del>	無無舞舞舞?[[[]{56}[[]木夕木]?[[[]{56}[]]木古木]
{57}	7	在存霭
{58}	切	斷留貿贸用斷盈荽觀鄧
{59}	何	舄僡殦廌橞潓舄舄寪糄縳襽殦舄駂
{60}	<b></b>	<b>龝</b> 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
{61}	戝	臧臧臧
{62}	華	<b>華華</b> 畢畢
{63}	舟	<del>敝</del> 組緒
{67}	Ħ	田高亭亮亳恫 <u>帽</u> 縠灀籉矗 <u>蠡</u> 諄鎘隝 厚駿豪臺章
{68}	里	黑黙廛堂邍爋顆瞖纒闥闉纒黨
{69}	並	<u> </u>
{70}	业	虚虡喋莁簏譔馳鷃蠽栞具 異

The numbered ideograph components in the above table significantly overlap the *CJK Supplementary Components for IDS Use* as listed in IRG N2225, which uses syntax similar to that of HTML character entities. For example,  $\{51\}$  ( $\mathbb{E}$ ) as shown in the table above is the same ideograph component as \$H6-03;

All existing ideograph components from a variety of sources will be collected, studied, and cataloged, which will serve as the basis for the formal proposal. A non-trivial number of ideograph components are already encoded as CJK Unified Ideographs, meaning that there is a precedent to continue to treat ideograph components as CJK Unified Ideographs.

Given the aggressive timeline, we hereby solicit help from experts in collecting, studying, and cataloging existing ideograph components. In addition, a new CJK Unified Ideographs exten-

sion block that includes sufficient unused code points for encoding future ideograph components should be considered. One suggestion is to use U+2FA20 through U+2FAFF for this new block, which provides 224 code points.

## Unicode Version 16.0 (2023)

With the four—or five—new IDCs and ideograph components encoded in Unicode Version 15.0, the target for Unicode Version 16.0 is to add the provisional *kIDS* property to the Unihan database, based on IDS.TXT. This IDS database, which is developed and maintained by Andrew West, has no copyright and is not encumbered by a license.

Delaying the addition of this new provisional property until Unicode Version 16.0 gives sufficient time for this IDS database to incorporate the new IDCs and new ideograph components, and also to make continued refinements to other IDSes. This should result in more robust property data.

That is all.